## Task Set 1 — Interpreting Algorithmic Functional Dependencies

### 1. Objective

The objective of Task Set 1 is to analyze the output of an FD discovery algorithm by focusing exclusively on structural properties of the discovered functional dependencies.
No semantic interpretation of attributes is performed.
The analysis is limited to:

- quantitative characteristics of the FD sets,
- structural patterns in determinants and dependents,
- identification of algorithmic red flags such as identifier-based dependencies, large determinants, and suspicious dependency patterns.

No FD discovery algorithms were run.

---

### 2. Methodology

For each dataset, we performed the following steps:

1. Read the provided list of minimal functional dependencies.
2. Computed:
   - the total number of functional dependencies,
   - the average size of the left-hand side (LHS),
   - the frequency of attributes appearing on the LHS and on the RHS.
3. Identified suspicious dependencies based only on structural criteria, namely:
   - identifier-based dependencies,
   - very large determinants,
   - repeated determinants determining many attributes,
   - unusually high numbers of FDs relative to dataset size.

No assumptions were made regarding the meaning or real-world plausibility of attributes.

---

### 3. Dataset-Level Analysis

#### 3.1 Iris

- Number of FDs: 4

- **Average LHS size: 3**

**Structural observations:**

- **All FDs have determinants of equal size.**
- **All FDs determine the same attribute.**
- **All descriptive attributes appear with similar frequency on the LHS.**

**Suspicious patterns:**

- **Determinants of size 3 for a dataset with only 4 descriptive attributes.**
- **Multiple equivalent determinants determining the same attribute, suggesting redundancy in the FD set.**

---

### 3.2 Balance-Scale

- **Number of FDs: 1**
- **Average LHS size: 4**

**Structural observations:**

- **A single FD with all descriptive attributes on the LHS.**
- **No smaller determinant exists.**

**Suspicious patterns:**

- **None from a structural perspective; the FD set is minimal and compact.**

---

### 3.3 Chess

- **Number of FDs: 1**
- **Average LHS size: 6**

**Structural observations:**

- **The only FD has a determinant consisting of all descriptive attributes.**
- **No attribute subset determines the RHS.**

**Suspicious patterns:**

- **Very large determinant, though no explosion or redundancy is observed.**

---

### 3.4 Nursery

- **Number of FDs: 1**
- **Average LHS size: 8**

**Structural observations:**

- **The determinant contains all descriptive attributes.**
- **No partial or alternative determinants are present.**

**Suspicious patterns:**

- **Extremely large determinant size relative to the number of attributes.**

---

### 3.5 Abalone

- **Number of FDs: 137**
- **Average LHS size: ≈ 4.2**

**Structural observations:**

- **Determinant sizes range from 3 to 6.**
- **The same determinants frequently appear across multiple FDs.**
- **Many determinants determine multiple different RHS attributes.**

**Suspicious patterns:**

- **Large number of FDs relative to dataset size.**
- **Presence of large determinants.**
- **Reuse of identical LHSs to determine many attributes, indicating quasi-identifier behavior.**

---

### 3.6 Breast-Cancer-Wisconsin

- **Number of FDs: 46**
- **Average LHS size: ≈ 2.8**

**Structural observations:**

- **Several FDs have a single attribute on the LHS.**
- **Other FDs involve moderate-sized determinants (3–5 attributes).**

**Suspicious patterns:**

- **Presence of identifier-based FDs (single attribute determining many others).**
- **Mixed presence of trivial and non-trivial determinants.**

---

### 3.7 Bridges

- **Number of FDs: 142**
- **Average LHS size: ≈ 3.6**

**Structural observations:**

- **FD count exceeds the number of rows.**
- **Many FDs involve the same attribute repeatedly on the LHS.**
- **Determinants of size 4 or more are frequent.**

**Suspicious patterns:**

- **FD explosion on a small dataset.**
- **Identifier-based dependencies.**
- **Large determinants suggesting overfitting.**

---

### 3.8 Echocardiogram

- **Number of FDs: 538**
- **Average LHS size: ≈ 4.9**

**Structural observations:**

- **Very high FD count relative to dataset size.**
- **Determinants frequently contain 4 or more attributes.**
- **High reuse of certain attributes on the LHS.**

**Suspicious patterns:**

- **Extreme FD explosion.**
- **Large determinants.**
- **Attribute combinations behaving as quasi-identifiers.**

---

### 3.9 Hepatitis

- Number of FDs: 8,250
- Average LHS size: ≈ 5.6

Structural observations:

- Extremely large FD set.
- Most FDs involve large determinants.
- Certain attributes dominate LHS frequency.

Suspicious patterns:

- Severe FD explosion.
- Large determinants across most dependencies.
- Strong quasi-identifier effects.

---

### 3.10 Adult (Special Case)

- Number of FDs: 78 (from summary table)
- Provided FD file: empty

Structural observations:

- No FDs could be analyzed from the distributed file.

Suspicious patterns:

- Discrepancy between reported FD count and empty FD file, indicating sensitivity of FD discovery to preprocessing and data representation.

---

### 3.11 Horse (Special Case)

- Number of FDs: 128,726 (from summary table)
- Provided FD file: empty

Structural observations:

- Extremely high reported FD count for a dataset with many attributes.

Suspicious patterns:

- **Expected FD explosion based on dimensionality.**
- **Practical limitations in FD extraction due to data sparsity and size.**

---

## 4. Cross-Dataset Structural Comparison

| **Dataset Type** | **Structural FD Behavior** |
|---|---|
| **Small rule-based datasets** | **Single FD with full determinant** |
| **Medium numeric datasets** | **Moderate FD explosion, large determinants** |
| **Small high-dimensional datasets** | **Extreme FD explosion** |
| **Large noisy datasets** | **Few or no exact FDs** |